媒体大数据服务平台设计与构建方法研究

摘 要：为了贯彻落实中央关于媒体融合发展的战略要求，需要积极应对互联网发展带来的传播格局调整和用户需求变化，努力构建与媒体发展趋势相适应、与建设新型一流媒体集团相适应的媒体大数据服务体系。通过汇聚内外媒体数据资源，紧密围绕媒体融合发展业务需求，构建大数据基础平台、大数据资源管理平台、大数据分析平台和大数据服务能力开放平台四大层级，逐步形成"数据整合、能力共享、应用创新"的媒体大数据工作体系。

关键词：媒体大数据；大数据平台；数据资源管理；数据分析；数据服务

中图分类号：G220.7　　文献标识码：A

文章编号：1671-0134（2018）09-064-03　　DOI：10.19483/j.cnki.11-4653/n.2018.09.025

文 / 陈珺

引言

根据中央关于推动传统媒体和新兴媒体融合发展的重要指示与要求，要强化互联网思维，坚持传统媒体和新兴媒体优势互补、一体发展，坚持以先进技术为支撑，内容建设为根本，推动传统媒体和新兴媒体在内容、渠道、平台、经营、管理等方面的深度融合。

为了贯彻落实中央关于媒体融合发展的战略要求，需要积极应对互联网发展带来的传播格局调整和用户需求变化，努力构建与媒体发展趋势相适应、与建设新型一流媒体集团相适应的媒体大数据服务体系。

1. 需求分析

随着传统媒体和新兴媒体融合发展的进一步深化，媒体企业在大数据资源整合、大数据资产管理、大数据分析挖掘能力建设以及数据服务开放共享等方面面临一系列问题，从而对技术系统的规划建设提出了更高的要求。

1.1 实现统一的大数据资源采集引进汇聚

媒体机构通过各种渠道采集和引进了大量外部数据，包括国内外互联网网站、数字报刊杂志、"两微一端"、社交媒体等。同时，媒体机构内部也产生各类稿件数据、产品数据、运营数据、用户行为数据等。如此众多的外部和内部数据分散存储在不同的部门和技术系统里，数据资源之间存在大量重复和冗余，数据关联关系没有打通，数据资源条块化分隔情况比较严重，数据资源共享和再利用能力较低。因此，需要整合机构现有大数据资源采集能力和引进能力，按需汇聚各类数据资源，实现数据资源的汇聚融合、开放共享和互联互通。

1.2 实现媒体大数据资产全生命周期管理

一个媒体大数据服务体系离不开高效的数据存储与计算基础平台，由于数据种类多、数据量大、计算处理效率不同，因此，对大数据存储与计算处理能力提出了更高的要求。需要基于互联网主流大数据平台技术架构，分层构建高效分布式媒体大数据存储与计算平台，能够实现 PB 量级的大数据存储和处理能力，并根据业务需要实现从实时到离线的不同数据处理效率。同时，需要实现对平台上所有媒体大数据资产的全生命周期管理，实现数据存储管理、标准管理、流程管理、质量管理和安全管理。

1.3 建设统一的大数据分析平台

现在，策划、采集、编写、发布、反馈等各类媒体业务环节越来越离不开大数据分析的支持，因此有必要进一步加强自然语言处理、数据挖掘、机器学习、数据可视化等智能信息处理技术创新，提升知识发现、大数据分析挖掘能力，助力提升策采编发供馈等各项媒体业务创新能力，提供满足业务需要的各类公共性媒体大数据分析服务。

1.4 提供开发共享的媒体大数据能力开放平台

通过制订统一的平台标准、数据标准、服务标准和管理标准，将媒体大数据平台形成的各项服务进行封装，实现这些服务的模块化和标准化，形成各类公共模型、工具和组件，提供面向各类媒体创新业务的公共性、基础性和开放共享的服务能力支撑。

2. 建设目标

基于互联网思维，汇聚内外媒体数据资源，围绕媒体融合发展业务需求，构建统一的媒体大数据服务平台，逐步形成"数据整合、能力共享、应用创新"的媒体大数据工作体系。

汇聚内外数据资源，形成媒体大数据服务体系；面向互联网思维，构建媒体大数据能力开放平台。

3. 总体架构设计

3.1 总体架构

媒体大数据服务体系从总体架构上可划分为大数据基础平台、大数据资源管理平台、大数据分析平台和大数据服务能力开放平台四个层级。

3.2 大数据基础平台

大数据基础平台是大数据存储管理以及分析计算运行的基础环境，包含大数据的基础运行环境搭建、资源任务调度管理、实时 / 离线计算支撑、结构化和非结构化

数据存储、数据检索、系统管理监控、数据访问的标准化 SQL 支持等功能。

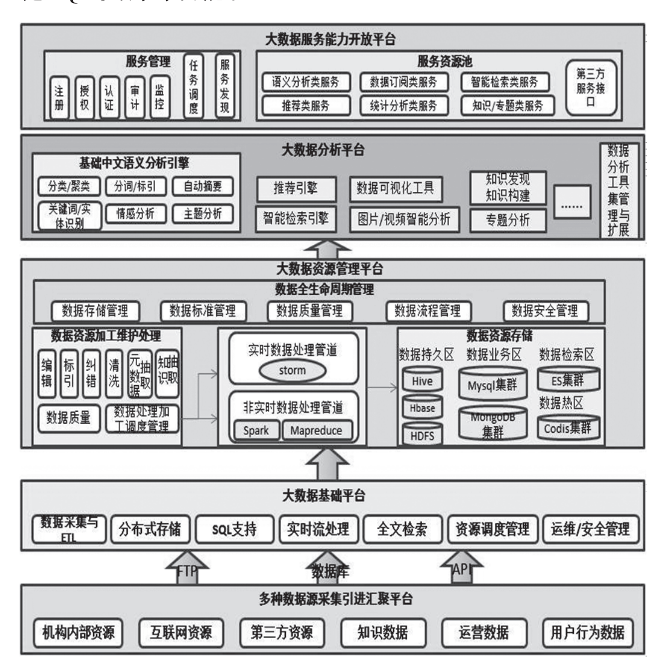


图 1　系统总体架构

可按需提供关系型数据库、列式数据库、分布式文件系统、分析型数据库、全文检索数据库、内存数据库等不同类型的数据存储资源。

可根据业务的使用场景以及数据自身的特点，提供合适的计算框架进行实时或离线的计算，完成分析功能。针对实时性要求不是很高的数据可使用 MapReduce 或 Hive 等进行非实时批处理，对响应时间要求比较高的业务场景可使用 Spark 做实时内存处理，对互联网流式数据则使用 Storm 或者 Spark Streaming 做实时流处理。

可针对不同的分析任务按需分配资源，进行资源管理调度，各分析任务之间相互不产生影响。

可针对分析算法或分析引擎，提供标准化的 SQL 支持。可提供大数据基础平台运行情况的管理和监控功能，便于系统管理员运维管理。

3.3 数据存储规划

考虑到数据类型、数据规模和数据增长量，采用分布式、高可用、可扩展的存储架构，实现对多来源数据、结构化数据和非结构化数据的统一存储规划设计，采用分区分域、分层分级、分库分表的设计理念，根据不同的数据类型合理选择数据存储组件，采用 MySQL、MongoDB、HBase、Hive、HDFS、ES、Codis 等多种数据库组件分别设计存储策略。

数据存储规划分为以下几个数据区：

3.3.1 实时汇集区

针对数据源层各种异构数据，我们需要采取多种数据接入方式，即可以使用传统的 FTP、Http、RPC 等接入方式，也能够支持 sqoop，flume 等这种以大数据为主的数据接入方式。针对互联网等大数据量数据，可以采用 Kafka 集群，充分发挥它的高吞吐量优势，主要用来临时保存互联网数据、行为数据、交易数据等实时数据。

3.3.2 大数据存储区

对接入的数据需要根据数据的特点和业务场景进行数据的存储，即支持传统型数据库也支持非传统型数据库。互联网数据可存储到 FastDFS、HDFS 这种分布式文件系统中，具备存储弹性，方便日后扩充，满足海量存储需求。对数据进行处理加工和分析后形成的结果数据包括内容数据和结构化数据，可以大对象存储在列式数据库 HBASE 中，并可通过 HIVE 对外提供 HSQL 标准服务，方便进一步进行非实时数据统计分析和数据挖掘使用。

3.3.3 数据业务区

核心业务数据、结构化数据、元数据的存储可采用 MySQL 关系型数据库集群进行保存，同时可利用 Mongodb 数据库的数组索引特性以及字段可扩展特性，存储数据的全部附加属性，做适当冗余，为数据服务提供高性能的读写能力。

3.3.4 数据检索区

利用像 ElasticSearch 这类全文检索数据库存储全部需检索数据，建立全文索引，实现大数据量的快速检索。

3.3.5 数据热区

为了实现快速访问建立数据热区，可利用 Codis 这类内存数据库存储需快速响应的热数据，提高系统整体数据访问效率。

3.4 大数据资源管理平台

大数据资源管理平台负责大数据资源汇聚、加工处理和数据全生命周期管理，是大数据服务体系建设中的核心环节之一。主要完成采集引进的多类异构数据资源的汇聚和出入库管理，数据的清洗加工处理、数据存储管理、数据标准管理、数据流程管理、数据质量管理和数据安全管理等功能。

3.4.1 数据资源汇聚和出入库管理

负责将机构内外不同数据来源的数据资源统一接入到数据平台中，支持文字、图片、音视频、文件、结构化数据、二进制文件等不同数据类型。制定相应的数据接口规范，采用统一的应用架构，以插件式开发和插件化使用的模式构建不同的数据流程任务，提供 FTP、消息队列、API 等不同接口方式，满足不同的业务流程和异构数据的出入库需求。数据在入库存储过程中需要先进行安全性检查与完整性校验，并进行初步数据清洗预处理，包括有效性检查和排重等，保证数据的可靠性，接入数据必须按照平台要求的数据格式规范统一进行转换后入库。并建立统一的数据汇聚出入库监控管理界面，能够支持任务各要素的灵活配置和定义，支持进行数据接入任务的监控和日常运维操作。

3.4.2 数据加工处理

负责对接入平台的各类型数据进行进一步的加工处理。对各类数据资源进行清洗、过滤、去重和转换等预处理工作；基于平台建立的一套数据标准，抽取元数据、关键词、实体信息等形成结构化描述信息；使用分词组

件对文本数据进行快速分词；使用分类技术对数据进行自动分类；对数据进行标引、加工、修改、纠错、删除等加工维护管理；建立搜索词典到文档数据的倒排索引表，根据词语在文档中的权重，为搜索词语生成相关索引文档表，结合分布式列存储与分层查询树技术，建立针对海量数据的全文检索和快速查询，支持更进一步的数据分析应用服务需求。

3.4.3 数据资源管理

负责对数据平台内所有数据资产进行全生命周期存储、管理和监控。对机构内数据、互联网数据等实现集中统一存储管理，对主数据、元数据和数据资源目录进行统一维护和管理，构建数据资源全景视图。实现数据标准管理、数据流程管理、数据质量管理和数据安全管理。

（1）数据质量和数据流程管理。为确保数据的完整性、规范性、一致性、准确性，提供统一数据处理流程和中间状态的调度、管理和监控，可以及时发现数据处理各个环节出现的问题和质量风险，对发现的异常进行报警。在数据入库环节，制定数据质量规则，对不符合质量规则的数据进行告警，并进行相应处理。管理员可以通过对规则的不断修改完善，不断提高入库数据的质量。

（2）元数据管理。元数据管理贯穿从数据采集引进、数据处理加工、数据分析和数据服务全流程环节，对各流程环节形成的数据的元数据进行标准定义、元数据生成和元数据管理维护，通过对元数据的管理形成数据服务平台统一的数据视图，为整个平台数据资源管理奠定基础。

（3）数据标准管理。制订融合媒体数据存管控相关标准规范，贯穿数据的采集引进、处理加工、存储管理、公共服务整个全生命周期和全工作流程，通过对标准的制订、维护和遵循，为平台实现全媒体数据的汇聚融合、统一管理和共享服务提供数据标准规范的指导。

3.5 大数据分析平台

大数据分析平台通过构建中文语义分析引擎、推荐引擎、智能检索引擎、知识推荐引擎、图片视频智能分析引擎、专题分析、数据可视化工具等媒体大数据公共性基础性智能处理模型工具组件，对平台中的大量数据资源进行深入分析，挖掘数据关系，构建知识网络，提升数据价值，助力策采编发供馈等各项媒体业务创新应用需求。

将这些算法模型进行模块化、服务化封装，针对媒体行业各类业务需求提供基础数据分析引擎和分析工具。通过标准化各类处理、分析、挖掘算法的输入输出参数和中间结果，提供标准化的服务接口，可以方便地读取、调用、管理和调优。

在系统运营过程中不断发现偏差点并进行有针对性的优化调整，支持对算法、模型、引擎的优化、新增和替换。同时，通过合理的计算架构的设计和相应的任务调度，保证算法运行在更高效的计算架构下。

提供对数据分析工具集的有效管理，建立信息库，统一存储和管理相关算法的代码、配置参数、调用接口规范、数据输入输出接口规范、文档说明、元数据等。

提供工具集对外交互界面，实现工具集的可视化、标准化和流程化使用和运行监控。提供工具集的扩展接口，可以根据业务需求将新增或第三方提供的数据分析算法工具纳入进来，统一调度和管理。

3.6 大数据服务能力开放平台

大数据服务能力开放平台负责将大数据平台的各类数据服务和分析服务进行封装并对外提供服务能力的开放和共享。大数据平台形成的服务能力有：数据订阅类服务、语义分析类服务、图片视频智能分析类服务、智能检索类服务、智能推荐类服务、知识类专题类服务、统计分析类服务、数据可视化等各种公共性服务能力。通过制订服务标准和管理标准，形成标准化服务模块和服务组件，提供标准化服务接口，为各类业务系统按需调用。同时，数据服务管理通过对服务的注册、认证、授权、审计、监控等管理功能，实现数据服务可管可控。

以面向服务的思想为核心理念，对服务进行高度解耦，构建细粒度、扁平化、低耦合的服务资源池，统一为上层应用提供功能和数据支撑。

将多源、异构数据以及关联数据等数据的获取方式进行接口化封装，实现基础数据服务化。

对数据分析计算层的数据处理分析算法、组件进行接口化封装，实现数据分析的服务化。

通过数据和应用封装技术，实现对数据的访问和操作按照一定粒度封装为独立的服务实体，尽可能屏蔽内部的细节，只提供标准化的交互接口，供各内部模块或者外部系统进行调用。交互接口形式包括 Open API、SDK、WEBService 等，实现自有业务应用支撑和开放共享服务。

建立服务管理平台作为服务注册和服务治理的管控中枢。媒体大数据服务平台向上层提供的服务通过服务管理平台进行统一管控，服务管理平台负责服务的注册、认证、授权、审计、监控等管理功能。媒

参考文献

[1] 周耀林，赵跃，Zhou Jiani. 大数据资源规划研究框架的构建 [J]. 图书情报知识，2017（4）：59-70.

[2] 梅剑平 . 大数据助力媒体融合——央视大数据平台技术与实践 [J]. 现代电视技术，2017（5）：100-104.

[3] 徐园，李伟忠 . 数据驱动新闻 智能重构媒体——浙报集团“媒立方”技术平台建设的实践与思考 [J]. 新闻与写作，2018（1）：97-101.

（作者单位：新华社技术局）